

Selecting the Optimal LHC Signatures for Distinguishing Models

Baris Altunkaynak

Department of Physics, Northeastern University, Boston, MA 02115

Abstract. An algorithm is developed which the goal of producing the most statistically significant signature list for distinguishing between two candidate models given a set of LHC observations.

Keywords: Collider physics, supersymmetry, statistical methods.

PACS: 12.60Jv, 14.80Ly

Recently in [1] we investigated the LHC inverse problem first rigorously posed by Arkani-Hamed et al. in [2] and showed that non-collider data can be used to remove the degeneracies observed in the collider data. In [3] we attacked the same problem in the context of determining the non-universality in the gaugino sector by using LHC signatures. In this short note we summarize the statistical methods we utilized to optimize the set of signatures in order to minimize the integrated luminosity required to resolve the degeneracies.

We define a chi-square like distance function between any two models A and B as the metric of the signature space which is very similar to the one used in [2] as

$$(\Delta S_{AB})^2 = \frac{1}{n} \sum_{i=1}^n \left[\frac{S_i^A - S_i^B}{\delta S_i^{AB}} \right]^2, \quad (1)$$

where S_i is the i^{th} counting signature and δS_i^{AB} is the uncertainty of the numerator, i.e. the difference between the signatures which we will assume to contain only statistical errors. We can identify any signature S_i with an “effective” cross section $\bar{\sigma}_i = S_i/L$ which includes the geometric cuts that are performed on the data, the detector efficiencies, etc. At large integrated luminosity this converges to an “exact” cross section $\sigma_i = \lim_{L \rightarrow \infty} \bar{\sigma}_i$. Rewriting the metric in terms of these effective cross sections gives us

$$(\Delta S_{AB})^2 = \frac{1}{n} \sum_{i=1}^n \left[\frac{\bar{\sigma}_i^A - \bar{\sigma}_i^B}{\sqrt{\bar{\sigma}_i^A/L_A + \bar{\sigma}_i^B/L_B}} \right]^2, \quad (2)$$

where L^A and L^B are the integrated luminosities that are used to compute the effective cross sections.

We can obtain the statistical properties of this metric by replacing each signature (or effective cross section) by a random variable following a normal distribution. After this randomization, the effective cross sections simply become

$$\bar{\sigma}_i = S_i^A/L_A = \sigma_i^A + \sqrt{\sigma_i^A/L_A} Z_A, \quad (3)$$

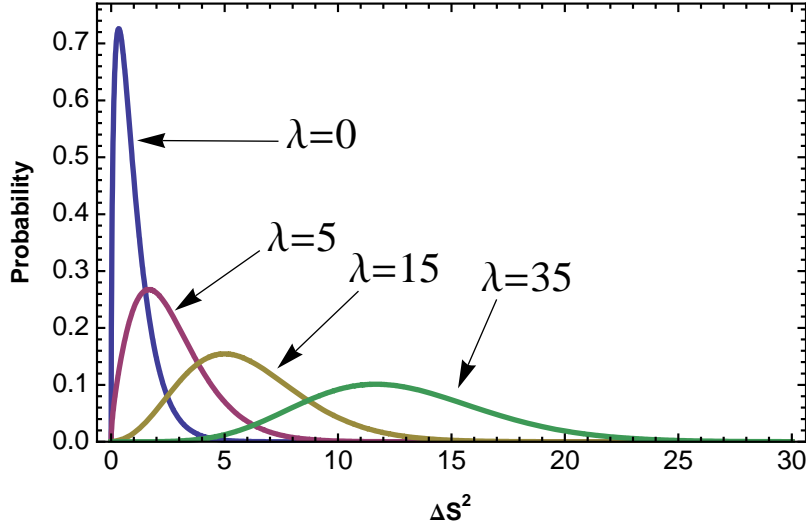


FIGURE 1. Plot of distribution in ΔS^2 values for $n = 3$ and various λ values.

with a similar expression for the model B .

Substituting (3) into (2) simply gives

$$\begin{aligned}
 (\Delta S_{AB})^2 &= \frac{1}{n} \sum_{i=1}^n \frac{\left[\sigma_i^A - \sigma_i^B + \sqrt{\frac{\sigma_i^A}{L_A} + \frac{\sigma_i^B}{L_B}} Z_i \right]^2}{\frac{\sigma_i^A}{L_A} + \frac{\sigma_i^B}{L_B} + \sqrt{\frac{1}{L_A^2} \frac{\sigma_i^A}{L_A} + \frac{1}{L_B^2} \frac{\sigma_i^B}{L_B}} Z'_i} \\
 &\approx \frac{1}{n} \sum_{i=1}^n \left[\frac{\sigma_i^A - \sigma_i^B}{\sqrt{\frac{\sigma_i^A}{L_A} + \frac{\sigma_i^B}{L_B}}} + Z''_i \right]^2, \tag{4}
 \end{aligned}$$

where Z_i, Z'_i and Z''_i are independent normally distributed random variables and assuming all Z''_i are independent, i.e. our n signatures are independent from each other, $(\Delta S_{AB})^2$ is itself a random variable having a non central chi-square distribution

$$P(\Delta S^2) = n \chi_{n,\lambda}^2(n\Delta S^2), \tag{5}$$

where λ is the non-centrality parameter which is given by

$$\lambda = \sum_{i=1}^n \frac{(\sigma_i^A - \sigma_i^B)^2}{\sigma_i^A/L_A + \sigma_i^B/L_B}. \tag{6}$$

Here, $\lambda = 0$ ($\neq 0$) corresponds to comparing a model to itself (to a different model) by using two sets of independent measurements. Figure 1 shows how the $(\Delta S_{AB})^2$ distribution favors larger values as λ increases. Since our goal is to tell apart two models, we want the possible $(\Delta S_{AB})^2$ values we will get from this comparison to be safely away from the possible values we get by comparing a model to itself, i.e. $\lambda = 0$ case. If we

TABLE 1. List of $\lambda_{\min}(n, p)$ values for various values of the parameters n and p .

n	Confidence Level p			
	0.95	0.975	0.99	0.999
1	12.99	17.65	24.03	40.71
2	15.44	20.55	27.41	44.99
3	17.17	22.60	29.83	48.10
4	18.57	24.27	31.79	50.66
5	19.78	25.71	33.50	52.88
6	20.86	26.99	35.02	54.88
7	21.84	28.16	36.41	56.71
8	22.74	29.25	37.69	58.40
9	23.59	30.26	38.89	59.99
10	24.39	31.21	40.02	61.48

quantify this safety condition as the requirement that $(100 \times p)\%$ of the distributions do not overlap, i.e. $(100 \times p)\%$ of the values we get by comparing the same model to itself are less than $(100 \times p)\%$ of the values we get by comparing two different models, we obtain the following equations

$$p = \int_0^\gamma n \chi_{n, \lambda=0}^2(n \Delta S^2) d(\Delta S^2) \rightarrow \Gamma\left(\frac{n}{2}, \frac{n}{2} \gamma\right) = \Gamma\left(\frac{n}{2}\right) (1 - p) \quad (7)$$

$$p = \int_\gamma^\infty n \chi_{n, \lambda_{\min}}^2(n \Delta S^2) d(\Delta S^2), \quad (8)$$

which can be solved numerically to compute a λ_{\min} value (see Table 1) for every number of signatures n and the non-overlap fraction (or confidence level) p . Here γ is the $(\Delta S)^2$ cut-off value for which $(100 \times p)\%$ of the values we get by comparing a model to itself is less than this value and this condition gives us Eqn (7) which can be solved numerically to compute γ . Then this γ value is used as the lower cut-off for the next equation which is solved again numerically to compute λ_{\min} .

The condition for two models to be distinguishable is simply $\lambda > \lambda_{\min}$. In this inequality λ_{\min} is just a numerically computed number which is independent of the physics involved in the collider experiment and all the physics is in λ which is a function of cross sections given by each signature.

Let us assume now that “model A” is the experimental data, which corresponds to an integrated luminosity of L^{exp} , and “model B” is the simulation with integrated luminosity $L^{\text{sim}} = qL^{\text{exp}}$. We might imagine that q can be arbitrarily large, limited only by computational resources. Let us make one final notational definition

$$R = \sum_{i=1}^N \frac{(\sigma_i^{\text{exp}} - \sigma_i^{\text{sim}})^2}{\sigma_i^{\text{exp}} + \frac{1}{q} \sigma_i^{\text{sim}}}, \quad (9)$$

then we can compute the minimum amount of luminosity required for two models to be distinguishable which is given by

$$L_{\min} = \frac{\lambda_{\min}(n, p)}{R}. \quad (10)$$

If the two models we want to compare are very similar in all the channels (signatures) we consider, then R will be small and L_{\min} will be large. If on the other hand the models are very different R will be large and L_{\min} will be small. This is of course what we expect, i.e. similar models require more integrated luminosity to distinguish.

Now the question is how to make L_{\min} as small as possible. We see from Table 1 that λ_{\min} increases as n increases and since R is a sum of positive quantities it increases with n as well. Therefore using more signatures does not necessarily help in distinguishing models and, moreover, the signature space is not big enough (or at least the relevant part of the signature space, see [2]) to allow multiple independent directions. It is easy to see the orthogonality of signatures such as number of events with 1 lepton and 2 leptons, but for more general cases, such as kinematic histograms which we can integrate between limits that are also optimized to increase distinguishability, we need to compute the correlation coefficient between different signatures a and b which is given by

$$\rho_{ab} = \frac{\text{cov}(a, b)}{\text{var}(a)\text{var}(b)} \approx \frac{\frac{1}{N} \sum_k [\bar{\sigma}_a^k - \sigma_a] [\bar{\sigma}_b^k - \sigma_b]}{\sqrt{\frac{1}{N} \sum_k [\bar{\sigma}_a^k - \sigma_a]^2} \sqrt{\frac{1}{N} \sum_k [\bar{\sigma}_b^k - \sigma_b]^2}} \quad \text{for large } N, \quad (11)$$

where the $\bar{\sigma}^k$ represent the individual results obtained from each of the N cross section measurements, labeled by the index k . This correlation matrix ρ_{ab} then can be used to determine the compatible observables, i.e. the ones which are not correlated with each other with more than some fixed threshold ε . This gives us the adjacency matrix of a graph which we define as

$$C_{ab} = \begin{cases} 1 & \text{if } |\rho_{ab}| \leq \varepsilon \\ 0 & \text{if } |\rho_{ab}| > \varepsilon. \end{cases} \quad (12)$$

Now finding the compatible observables is equivalent to finding all the complete subgraphs (or ‘clique’) of that graph which is a well known problem in graph theory. All these complete subgraphs give us an L_{\min} value and obviously the one giving the minimum of all these graphs contains the list of the signatures we want to combine together.

ACKNOWLEDGMENTS

The author is supported by National Science Foundation Grant PHY-0653587.

REFERENCES

1. B. Altunkaynak, M. Holmes and B. D. Nelson, JHEP **0810**, 013 (2008) [arXiv:0804.2899 [hep-ph]].
2. N. Arkani-Hamed, G. L. Kane, J. Thaler and L. T. Wang, JHEP **0608**, 070 (2006) [arXiv:hep-ph/0512190].
3. B. Altunkaynak, P. Grajek, M. Holmes, G. Kane and B. D. Nelson, JHEP **0904**, 114 (2009) [arXiv:0901.1145 [hep-ph]].